BIOMETRICS 50, 350–357
June 1994

Flexible Maximum Likelihood Methods for Assessing Joint Effects in Case–Control Studies with Complex Sampling

S. Wacholder

Biostatistics Branch,
National Cancer Institute,
Executive Plaza North, Room 403,
6130 Executive Boulevard,
Rockville, Maryland 20852, U.S.A.

and

C. R. Weinberg

Statistics and Biomathematics Branch,
National Institute of Environmental Health Sciences,
M.D. B3-02,
P.O. Box 12233,
Research Triangle Park, North Carolina 27709, U.S.A.

SUMMARY

Case–control studies can often be made more efficient by using frequency matching, randomized recruitment, stratified sampling, or two-stage sampling. These designs share two common features: (1) some "first-stage" variables are ascertained for all study subjects, while complete variable ascertainment is carried out for only a selected subsample, and (2) the subsampling of subjects for "second-stage" variable ascertainment depends jointly on their disease status and their observed first-stage variables. Because first-stage variables alter the subsampling fractions, standard analyses require a multiplicative specification of any joint effects of a second- and a first-stage variable. We show that by making use of missing data methods, maximum likelihood estimates can be obtained for risk parameters of interest, even those characterizing interactions between first- and second-stage variables. Joint effects can thus be modelled flexibly, with allowance for both additive and multiplicative models. Preliminary data from a case–control study of lung cancer as related to age, sex, and smoking provide an example, leading to the suggestion that the combined effect of age and smoking is multiplicative.

1. Introduction

The case–control study can offer marked efficiency advantages over a prospective study when the disease outcome of interest is rare (Breslow and Day, 1980). This design, also known as "choice-based sampling" in econometrics, calls for separate sampling among cases and among members of the same population who are free of the disease of interest. Four "complex sampling" strategies have been developed in which the sampling depends not only on disease status but also on other factors. These are: frequency matching, randomized recruitment, stratified sampling, and two-stage sampling. These designs have two important features in common. First, all are two-stage in the sense that some variables (to be referred to as "first-stage variables") are ascertained for all participants, and others ("second-stage variables") are gathered only for random subsamples. Second, the selection of subjects for second-stage variable ascertainment depends on their disease status and their observed first-stage variables.

While these designs can improve efficiency, they have imposed unacceptable constraints on the investigator who is interested either in main effects for first-stage variables or in interactions

Key words: Biased sampling; EM algorithm; Interaction; Matching; Randomized recruitment; Synergism; Two-stage sampling.

Material may be protected by copyright law (Title 17, U.S. Code)

between first- and second-stage variables: the usual analyses require a multiplicative (logistic) model (Breslow and Day, 1980). In this paper we propose a general maximum likelihood approach to analysis of case–control studies with complex sampling. Our approach allows efficient inference regarding both first- and second-stage variables, and frees the investigator from the need to assume a multiplicative model. We begin by briefly reviewing the four sampling paradigms.

Frequency matching can enhance efficiency in a case–control study when there is a variable that is readily ascertained and also strongly related to risk of the disease under study (Kleinbaum, Kupper, and Morgenstern, 1982). Such a variable can be determined during the first stage, for example, based on a brief, preliminary screening interview. Age and sex are typical matching factors. In frequency matching, the empirical distributions of the matching factors are constrained to be the same for cases and controls, by requiring the number of controls selected for a given stratum to be a fixed multiple of the number of cases observed in that stratum. In standard analyses (Breslow and Day, 1980), only subjects in the second-stage subsample are included in the analysis, whereas those with only first-stage data, i.e., those who were screened but not enrolled for full participation in the study, are ignored. Although it remains a popular strategy, matching has been criticized because of the associated practical difficulties and because it is presumed that one can neither estimate the effects of matching factors nor model the interaction between the exposure of interest and a matching variable in any way but multiplicatively (Thomas and Greenland, 1985).

Weinberg and Wacholder (1990; see also Weinberg and Sandler (1991)) proposed a different design, termed "randomized recruitment" (originally, "biased sampling"), that overcomes some of the practical problems associated with frequency matching and also allows for a more flexible analysis. In this approach, the selection of each potential participant for further study is determined randomly, based on Bernoulli sampling probabilities which are set by the investigator and may depend on both disease status and the person's first-stage information ascertained at screening. This design is being used in a case–control study of lung cancer and exposure to radon gas. Randomized recruitment allows us to oversample nonsmoking cases and match controls to cases on cigarette smoking, thereby enhancing efficiency, without the need to assume that the joint effects of smoking and radon exposure are multiplicative (Weinberg and Sandler, 1991).

Other efficient designs have been proposed for special circumstances. Fears and Brown (1986; see also Breslow and Zhao, 1988) considered stratified sampling in a case–control study. In their example, information on disease status was known for everyone in each of several cities. Thus city of residence was the first-stage variable. Covariate information (second-stage) was then obtained on within-city samples of cases and controls, but the variable of primary interest was an exposure level that was specific to each city. They proposed a method for logistic modelling that takes the unequal sampling fractions into account and estimates the effect of the stratum-associated exposure variable.

White (1982; see also Walker, 1982) and Breslow and Cain (1988) considered a design that they termed "two-stage" sampling. Here information on an exposure was assumed known for a large sample of cases and controls. White showed that efficiency can be enhanced by limiting ascertainment of covariate information to subsamples within the disease-by-exposure categories. Breslow and Cain described a maximum conditional pseudo-likelihood approach to logistic analysis for this two-stage design.

For each of these four designs second-stage covariates are missing by design for some individuals and can be considered to be "missing at random" in the sense of Little and Rubin (1987, p. 90). The purpose of this paper is to propose a unified maximum likelihood approach to analysis that fully exploits the first-stage data and is easy to implement using available software, such as Generalized Interactive Linear Modelling (GLIM, Baker and Nelder, 1978). We shall show that one can estimate effects of first-stage factors and fit nonmultiplicative models for the interaction between a first- and a second-stage factor, provided one uses the data available for potential subjects who were screened but not enrolled for the second stage of the study.

The paper is organized as follows. The second section sets up the notation and appropriate likelihoods. The third describes the implementation of the expectation–maximization (EM) algorithm as described by Dempster, Laird, and Rubin (1977). The fourth applies the method to preliminary data available on age, sex, and smoking history for subjects screened for participation in a case–control study of lung cancer and radon.

2. The Observed-Data and Complete-Data Likelihoods

To simplify the notation and the presentation, we assume that there are two categorical covariate vectors, $\mathbf{Z}_1$, and $\mathbf{Z}_2$, corresponding to first- and second-stage data, respectively. We assume that subjects are sampled randomly for ascertainment of $\mathbf{Z}_2$ within each stratum defined by disease status and $\mathbf{Z}_1$, although the subsampling fractions may vary across disease status and $\mathbf{Z}_1$. The first-stage

Material may be protected by copyright law (Title 17, U.S. Code)

between first- and second-stage variables: the usual analyses require a multiplicative (logistic) model (Breslow and Day, 1980). In this paper we propose a general maximum likelihood approach to analysis of case–control studies with complex sampling. Our approach allows efficient inference regarding both first- and second-stage variables, and frees the investigator from the need to assume a multiplicative model. We begin by briefly reviewing the four sampling paradigms.

Frequency matching can enhance efficiency in a case–control study when there is a variable that is readily ascertained and also strongly related to risk of the disease under study (Kleinbaum, Kupper, and Morgenstern, 1982). Such a variable can be determined during the first stage, for example, based on a brief, preliminary screening interview. Age and sex are typical matching factors. In frequency matching, the empirical distributions of the matching factors are constrained to be the same for cases and controls, by requiring the number of controls selected for a given stratum to be a fixed multiple of the number of cases observed in that stratum. In standard analyses (Breslow and Day, 1980), only subjects in the second-stage subsample are included in the analysis, whereas those with only first-stage data, i.e., those who were screened but not enrolled for full participation in the study, are ignored. Although it remains a popular strategy, matching has been criticized because of the associated practical difficulties and because it is presumed that one can neither estimate the effects of matching factors nor model the interaction between the exposure of interest and a matching variable in any way but multiplicatively (Thomas and Greenland, 1985).

Weinberg and Wacholder (1990; see also Weinberg and Sandler (1991)) proposed a different design, termed "randomized recruitment" (originally, "biased sampling"), that overcomes some of the practical problems associated with frequency matching and also allows for a more flexible analysis. In this approach, the selection of each potential participant for further study is determined randomly, based on Bernoulli sampling probabilities which are set by the investigator and may depend on both disease status and the person's first-stage information ascertained at screening. This design is being used in a case–control study of lung cancer and exposure to radon gas. Randomized recruitment allows us to oversample nonsmoking cases and match controls to cases on cigarette smoking, thereby enhancing efficiency, without the need to assume that the joint effects of smoking and radon exposure are multiplicative (Weinberg and Sandler, 1991).

Other efficient designs have been proposed for special circumstances. Fears and Brown (1986; see also Breslow and Zhao, 1988) considered stratified sampling in a case–control study. In their example, information on disease status was known for everyone in each of several cities. Thus city of residence was the first-stage variable. Covariate information (second-stage) was then obtained on within-city samples of cases and controls, but the variable of primary interest was an exposure level that was specific to each city. They proposed a method for logistic modelling that takes the unequal sampling fractions into account and estimates the effect of the stratum-associated exposure variable.

White (1982; see also Walker, 1982) and Breslow and Cain (1988) considered a design that they termed "two-stage" sampling. Here information on an exposure was assumed known for a large sample of cases and controls. White showed that efficiency can be enhanced by limiting ascertainment of covariate information to subsamples within the disease-by-exposure categories. Breslow and Cain described a maximum conditional pseudo-likelihood approach to logistic analysis for this two-stage design.

For each of these four designs second-stage covariates are missing by design for some individuals and can be considered to be "missing at random" in the sense of Little and Rubin (1987, p. 90). The purpose of this paper is to propose a unified maximum likelihood approach to analysis that fully exploits the first-stage data and is easy to implement using available software, such as Generalized Interactive Linear Modelling (GLIM, Baker and Nelder, 1978). We shall show that one can estimate effects of first-stage factors and fit nonmultiplicative models for the interaction between a first- and a second-stage factor, provided one uses the data available for potential subjects who were screened but not enrolled for the second stage of the study.

The paper is organized as follows. The second section sets up the notation and appropriate likelihoods. The third describes the implementation of the expectation–maximization (EM) algorithm as described by Dempster, Laird, and Rubin (1977). The fourth applies the method to preliminary data available on age, sex, and smoking history for subjects screened for participation in a case–control study of lung cancer and radon.

2. The Observed-Data and Complete-Data Likelihoods

To simplify the notation and the presentation, we assume that there are two categorical covariate vectors, $\mathbf{Z}_1$, and $\mathbf{Z}_2$, corresponding to first- and second-stage data, respectively. We assume that subjects are sampled randomly for ascertainment of $\mathbf{Z}_2$ within each stratum defined by disease status and $\mathbf{Z}_1$, although the subsampling fractions may vary across disease status and $\mathbf{Z}_1$. The first-stage

Material may be protected by copyright law (Title 17, U.S. Code)

variable, $\mathbf{Z}_1$, could be stratum (in stratified sampling), or exposure level (in two-stage sampling), or cross-categorized level of some set of matching variables (in frequency matching or randomized recruitment). Let n_{ijk} denote the number of people with complete covariate status known, who are at level i of disease, D, level j of $\mathbf{Z}_1$ (where j indexes the vector among all possible realizations), and level k of $\mathbf{Z}_2$. For ease of notation we shall adopt the convention $\mathbf{Z}_i = j$ to mean that the vector $\mathbf{Z}_i$ is at its jth level. We employ dot notation, so that $n_{ij.}$ denotes summation over k. The uppercase N will refer to all subjects sampled, whether or not they had $\mathbf{Z}_2$ ascertained. Thus N_{ij} is the number of subjects with disease category i who were at the jth level of the first-stage variable. Note that $N_{i.}$ is larger than $n_{i..}$ with this notation.

2.1 *The Likelihood for Two-Stage, Frequency-Matched, or Randomized Recruitment Designs*

The observed-data likelihoods for two-stage, frequency-matched, or randomized recruitment designs all take the following form:

$$\prod_j \left\{ \left[\prod_i \Pr(\mathbf{Z}_1 = j | D = i)^{N_{ij}} \right] \left[\prod_{i,k} \Pr(\mathbf{Z}_2 = k | D = i,\ \mathbf{Z}_1 = j)^{n_{ijk}} \right] \right\}. \tag{1}$$

The leftmost factor corresponds to the first-stage data gathered for all participants, which is simply multinomial within each disease category. The rightmost factor corresponds to second-stage data, again multinomial within each D-by-$\mathbf{Z}_1$ category.

If the data were complete, so that the level of $\mathbf{Z}_2$ were known for all study subjects, then the likelihood (1) would reduce to the retrospective form of the case–control likelihood:

$$\prod_{i,j,k} \Pr(\mathbf{Z}_1 = j,\ \mathbf{Z}_2 = k | D = i)^{h_{ijk}}, \tag{1.1}$$

where h_{ijk} denotes the unknown (hypothetical) number with $D = i$, $\mathbf{Z}_1 = j$, and $\mathbf{Z}_2 = k$.

We assume the general multiplicative-intercept model for $\Pr(D = i | \mathbf{Z}_1 = j, \mathbf{Z}_2 = k)$, as described by Hsieh, Manski, and McFadden (1985). Under this model $\Pr(D = 1 | \mathbf{Z}_1, \mathbf{Z}_2)/\Pr(D = 0 | \mathbf{Z}_1, \mathbf{Z}_2)$, the disease odds, is $rf((\mathbf{Z}_1, \mathbf{Z}_2); \beta)$ for some positive scalar r, function f, and vector β. Inference about β can be validly carried out [Anderson (1979), as extended by Weinberg and Wacholder (1993)] by restricting attention to the prospective form:

$$\prod_{i,j,k} \Pr(D = i | \mathbf{Z}_1 = j,\ \mathbf{Z}_2 = k)^{h_{ijk}}. \tag{1.2}$$

The sufficient statistics are the frequencies, h_{ijk}, and the expectation–maximization (EM) algorithm will be applied in Section 3 to maximize the observed-data likelihood (1).

2.2. *The Likelihood for a Stratified Sampling Design*

The observed-data likelihood is slightly different for the stratified sampling design because there is a complete enumeration of D status within each level of $\mathbf{Z}_1$. As before, $\mathbf{Z}_2$ status is ascertained for random subsamples conditional on D and $\mathbf{Z}_1$. The likelihood is

$$\prod_j \left\{ \left[\prod_i \Pr(D = i | \mathbf{Z}_1 = j)^{N_{ij}} \right] \left[\prod_{i,k} \Pr(\mathbf{Z}_2 = k | D = i,\ \mathbf{Z}_1 = j)^{n_{ijk}} \right] \right\}. \tag{2}$$

With complete data, the likelihood (2) would become

$$\prod_{i,j,k} \Pr(\mathbf{Z}_2 = k,\ D = i | \mathbf{Z}_1 = j)^{h_{ijk}}$$

or, rewritten in prospective form:

$$\left[\prod_{i,j,k} \Pr(D = i | \mathbf{Z}_1 = j,\ \mathbf{Z}_2 = k)^{h_{ijk}} \right] \left[\prod_{j,k} \Pr(\mathbf{Z}_2 = k | \mathbf{Z}_1 = j)^{h_{.jk}} \right]. \tag{2.1}$$

The nuisance parameters, specifying the distribution of $\mathbf{Z}_2$ conditional on $\mathbf{Z}_1$, can be ignored and inference carried out based on the prospective factor alone, as if the populations had been studied prospec

Material may be protected by copyright law (Title 17, U.S. Code)

3. Application of the EM Algorithm

For each of the four complex-sampling scenarios, the observed-data likelihood can be maximized by considering the resulting data structure in the context of a missing data problem, as described by Dempster et al. (1977), where data are missing at random. In each design the level of $\mathbf{Z}_2$ is missing for people who provide only the marginal, first-stage data.

The expectation–maximization (EM) algorithm can be used to maximize (1) and (2). In the E step the complete-data sufficient statistics, h_{ijk}, are replaced by their expectations conditional on the observed data and the current estimates of the parameters of the risk model. To calculate these expectations, first define $m_{ij} = N_{ij} - n_{ij.}$, i.e., the number of subjects with $D = i$, and $\mathbf{Z}_1 = j$ who are missing $\mathbf{Z}_2$. Then set

$$\hat{h}_{ijk} = E(h_{ijk}|n_{ijk}, m_{ij}, \hat{\beta}) = n_{ijk} + m_{ij}\widehat{\Pr}(\mathbf{Z}_2 = k|D = i, \mathbf{Z}_1 = j)$$

$$= n_{ijk} + m_{ij} \frac{\widehat{\Pr}(D = i|\mathbf{Z}_1 = j, \mathbf{Z}_2 = k)\widehat{\Pr}(\mathbf{Z}_2 = k|\mathbf{Z}_1 = j)}{\Sigma_r \widehat{\Pr}(D = i|\mathbf{Z}_1 = j, \mathbf{Z}_2 = r)\widehat{\Pr}(\mathbf{Z}_2 = r|\mathbf{Z}_1 = j)}$$

$$= n_{ijk} + m_{ij} \frac{\widehat{\Pr}(D = i|\mathbf{Z}_1 = j, \mathbf{Z}_2 = k)\dfrac{\hat{h}_{.jk}}{\hat{h}_{.j.}}}{\Sigma_r \widehat{\Pr}(D = i|\mathbf{Z}_1 = j, \mathbf{Z}_2 = r)\dfrac{\hat{h}_{.jr}}{\hat{h}_{.j.}}}$$

$$= n_{ijk} + m_{ij} \frac{F_{ijk}}{\Sigma_r F_{ijr}},$$

where F_{ijk} denotes the current estimate of the fitted value for the ijkth cell. Thus the E step reestimates the complete data as the sum of the observed frequency, n_{ijk}, and a fraction of the missing data, m_{ij}, where the fraction of the missing-data people assigned to a given level of $\mathbf{Z}_2$ is the relative magnitude of the fitted value for that cell.

The maximization (M) step then maximizes the likelihood, as if the frequencies estimated in the E step were the complete data from this case–control study. Standard software can be used. Alternation of the E and the M steps leads to convergence to the maximum of the observed-data likelihood (Dempster et al., 1977). Of course, the log-likelihoods based on the estimated data are incorrect, and routines must be written for computing the likelihood appropriate to the observed data.

4. Example: Age, Sex, Smoking, and Lung Cancer

Randomized recruitment is being used in an ongoing case–control study of residential exposure to the radioactive gas, radon, and lung cancer, being carried out at two locations: Utah/southern Idaho and Connecticut. The objectives of the study are to determine whether residential exposure to radon increases the risk of lung cancer and to characterize the joint effects of radon and cigarette smoking in causing the disease. Radon dosimetry requires measurements in current and former residences, which is much more costly than asking about the known strong risk factors, age and smoking status, in an interview. Randomized recruitment should greatly enhance efficiency in this setting (Weinberg and Sandler, 1991).

We begin by briefly describing the sampling process. Controls younger than 65 are identified through random digit dialing (RDD) (see Wacholder, 1992), and those over 65 are sampled at random from Health Care Financing Administration (HCFA) lists. (The HCFA maintains lists of residents for Medicare compensation, which are considered to be virtually complete for those over 65.) After age and sex have been ascertained, a potential control is recruited for further participation if a generated uniform random number falls below the chosen sampling probability. See Weinberg and Sandler (1991) for tables of probabilities. This achieves approximate matching of controls to cases in their age/sex distribution. Potential controls who pass the initial age/sex randomization test are next assigned, based on a brief interview, to one of four smoking categories, depending on their behavior 10 years prior to interview, as follows: never-smoker, ex-smoker, light smoker (fewer than 20 cigarettes a day), or heavy smoker (20 or more cigarettes a day). A second randomization is applied to cases and controls based on smoking status, in order to oversample nonsmoking cases and match controls to cases in their smoking histories. Radon dosimetry is attempted only for subjects who pass both randomization tests.

Preliminary screening results from Utah/southern Idaho will be used to illustrate the method. We

Material may be protected by copyright law (Title 17, U.S. Code)

have age and sex data for a large group of potential participants but have only smoking status for a much reduced subsample. No radon data are available yet. We shall focus on the joint effects of age and smoking. The final analysis to be performed will be similar, but will focus instead on the combined effects of radon (a third-stage variable) and smoking (a second-stage variable).

Tables 1a and 1b show the data, including the marginal data shown in the rightmost column labelled "undetermined" for subjects for whom we know age and sex, but not smoking history. Note that many controls are missing data by design.

Table 1
Age and smoking status for cases and controls

		Smoking				
Age	Cancer	Never	Ex-smoker	Light	Heavy	Undetermined
a. Females						
40–59	Yes	7	3	15	34	8
	No	320	26	31	37	2,324
60–64	Yes	6	0	10	22	7
	No	135	10	11	17	295
65–69	Yes	7	6	12	34	15
	No	228	26	12	18	863
70–74	Yes	6	3	11	28	7
	No	214	18	12	16	859
75–79	Yes	8	7	8	7	10
	No	156	11	11	0	660
b. Males						
40–59	Yes	6	2	6	73	13
	No	385	127	47	121	1,909
60–64	Yes	2	7	7	65	16
	No	210	99	22	67	39
65–69	Yes	7	18	10	63	25
	No	240	136	31	75	528
70–74	Yes	8	25	16	54	28
	No	289	164	26	62	362
75–79	Yes	7	17	11	25	17
	No	195	119	11	25	287

The analysis was carried out for three different models, each assuming binomial errors: the linear logistic, the odds-linear, and the purely additive. We assume the following generalized linear model form (McCullagh and Nelder, 1989):

$$g[\Pr(D = 1|\text{age group} = i,\ \text{sex} = j,\ \text{smoking category} = k)] = \mu + \alpha_i + \beta_j + \gamma_k,$$

so that age and smoking are modelled freely as unordered categories. The function g is sometimes referred to as the "link" function, since it links the expected value of the outcome of interest to a linear function of predictors. For the logistic model, the function g is the logit link, $g(x) = \ln[x/(1 - x)]$, for the odds-linear model $g(x) = x/(1 - x)$, and for the purely additive model g is the identity, $g(x) = x$. The linear logistic model is a standard component of GLIM (and other software), and simple macros are available for fitting the other two links (Weinberg and Sandler, 1991; Wacholder, 1986).

The results of various model fits and various links are shown in Table 2, which gives scaled "deviances" ($-2\log$(maximized likelihood)) to allow likelihood ratio testing. To avoid parameters estimated at infinity, two 0 frequency values were replaced with .1. All results obtained by EM were confirmed using the Newton–Raphson algorithm. The deviances have been corrected by subtracting a constant from each, to set the value at full saturation to 0.0.

The fits of the fully saturated models are the same for the three link specifications, since the fitted and observed numbers of cases are equal in each cell. The differences between the deviances for the fully saturated models and those for the models that include only age and sex are more than 1,000 for each link specification. Under the correct model, this should be approximately chi-squared with 30 degrees of freedom if smoking were unrelated to risk of lung cancer. Clearly we cannot ignore smoking. Note, by contrast, that adding sex to the model produces negligible improvement in fit.

Material may be protected by copyright law (Title 17, U.S. Code)

Table 2
Scaled deviances by link and model specifications

Factors in model	Number of parameters	Link for model		
		Logit	Odds	Identity
All factors including interactions	40	0.0	0.0	0.0
Age Smoking	8	43.7	208.1	217.1
Age Sex Smoking	9	40.7	206.5	215.6
Age Sex	6	1,103.7	1,148.9	1,143.3
Age Sex Age-by-Sex	10	1,095.7	1,095.7	1,095.7

The logistic model suffers little loss of fit when reduced to just main effects for age and smoking (the likelihood ratio statistic is 43.7, with 32 degrees of freedom). By contrast, the deviances for the other two links are both greater than 200, indicating a very poor fit. Thus the additive models demand a more complex specification, involving interaction terms.

Parsimony thus argues strongly in favor of the eight-parameter multiplicative model for describing the joint effect of smoking and age. The goodness of fit for this model was confirmed by noting that the observed numbers of cases and controls agree closely with the resulting fitted numbers (not shown), both within age/sex/smoking strata and for marginal totals within age/sex strata. By contrast, when either of the other links is applied, the corresponding differences are huge.

Since the multiplicative model gives the best fit for these data, we can compare results obtained using maximum likelihood with those based on the maximum conditional pseudo-likelihood method described by Breslow and Cain (1988) for two-stage case–control data, which also assumes a logistic model. Results are shown in Table 3. Both the parameter estimates and their estimated standard errors are nearly identical for the two methods. The relative risks for ex-, light, and heavy smoking are 5.8, 21.5, and 49.4, respectively.

Table 3
Maximum likelihood estimates of coefficients under a logistic model

Parameter	Logistic ML[a]	SE_{ML}[b]	Logistic MCPL[c]	SE_{MCPL}[d]
Age 40–59	−5.73	.16	−5.71	.15
Age 60–64	−4.08	.17	−4.06	.16
Age 65–69	−4.47	.15	−4.46	.15
Age 70–74	−4.26	.15	−4.25	.14
Age 75–79	−4.01	.16	−3.99	.15
Ex-smoking	1.77	.17	1.75	.17
Light smoking	3.07	.18	3.06	.18
Heavy smoking	3.90	.15	3.88	.15

[a] By maximum likelihood.
[b] Standard errors based on the observed information matrix.
[c] By maximum conditional pseudo-likelihood estimation.
[d] Based on the consistent estimator given by Breslow and Cain (1988, Proposition 2).

There remains the inevitable uncertainty in the absolute risk estimates based on case–control data alone, because of the unknown relative probability of being identified for study for cases versus controls (Weinberg and Wacholder, 1993); nonetheless relative risks can be estimated from this model, as in a standard analysis. Thus we can estimate effects associated with "matching" factors. For example, although age was a matching factor in the study of lung cancer, we can estimate the relative risk for lung cancer for those 75–79 compared to those 65–69, as exp(−4.01 + 4.26) = 1.3 (see Table 3). (Comparisons for age groups above and below 65 could *not*, however, be made using

Material may be protected by copyright law (Title 17, U.S. Code)

these data, since the probabilities of being identified for possible recruitment as controls are different for RDD and HCFA.) This kind of comparison would, of course, be most interesting under the two-stage paradigm where the exposure of interest to the study is a first-stage variable.

5. Discussion

When a case–control study has employed complex sampling, maximum likelihood analysis can be carried out despite incomplete data by applying the EM algorithm. Using this approach, we have shown that one can compare the fits of additive and multiplicative models (or more general models) to decide which best characterizes the interaction between the first-stage (e.g., matching) and second-stage variables. Contrary to classical assumption (e.g., Thomas and Greenland, 1985), we have shown that relative risks for first-stage variables can also be estimated. Thus, provided full use is made of marginal data, one can both estimate relative risks associated with matching variables and fit general models that allow for nonmultiplicative interactions.

In the example, the most parsimonious model for the combined effect of age and smoking on lung cancer risk was multiplicative, consistent with the possibility that the two factors operate at different stages in a single multistage carcinogenic process (Siemiatycki and Thomas, 1982). Additivity would have suggested biologically independent mechanisms. In this context, age could be considered a proxy for cumulative exposure to lung carcinogens other than cigarette smoke. The analysis will become more interesting when the exposures being analyzed are smoking and radon, and the timing of exposure can be taken into account. Knowing the way risks from these exposures combine will facilitate estimation of the public health impact of residential radon exposure and the potential benefits to be gained from expensive amelioration programs.

Other methods have been proposed for two-stage sampling. Scott and Wild (1991) used maximum likelihood in the context of stratified sampling, where the logistic model was assumed, but employed Fisher scoring to maximize the observed data likelihood. They showed that maximum likelihood can be much more efficient than the method of Fears and Brown (1986), but can be computationally cumbersome. Flanders and Greenland (1991) adapted the pseudo-likelihood method developed by Kalbfleisch and Lawless (1988) to nested case–control studies based on stratified two-stage sampling.

To simplify the description of the method, we assumed there were exactly two stages to the sampling, namely a screening stage (or enumeration stage in the case of stratified sampling, or record-retrieval stage in the case of two-stage sampling) followed by a stage where there is full variable ascertainment for subsamples. The method has a natural generalization to more than two levels of nesting on variable completeness. In the radon study, there will be three levels of data completeness: some subjects will have only age and sex known, a subsample will have smoking status known as well, and an even smaller subsample will be eligible to have radon measurements made on their current and former residences. In the E step of the EM algorithm, one first goes from the first stage to the second stage (in the example, estimating the age/sex/smoking/disease frequencies), ignoring the third stage, as we have done. Next the second-stage estimated frequencies for people with missing third-stage data are distributed among the levels of the third-stage variable (e.g., cumulative radon levels) according to the relative magnitudes of the fitted values. The M step, as before, then fits the appropriate model to the estimated frequencies from the E step. In this way, the three-factor (or more) likelihood analogous to (1) or (2) can be maximized.

Finally, the method described will not be universally applicable following a complex-sampling design. The approach requires additional assumptions, for example, if there are eligibility requirements that have not been imposed on all subjects. Suppose, for example, that some subjects are randomized for recruitment based on existing records, without first being contacted to confirm eligibility. In this situation it might not be correct to assume data are missing at random, because ineligible people have been included in the marginals. Here the investigator should consider applying the offset-adjusted method described by Weinberg and Wacholder (1990). That method would still be valid, since with Bernoulli sampling the same probabilities have been imposed impartially on the eligible and the ineligible.

Acknowledgements

We wish to thank Drs M. Gail, G. Dinse, and D. Sandler for helpful comments on the manuscript, and M. Bishop and R. Gress for technical assistance in supplying the data.

Résumé

Les études cas–témoin peuvent souvent être rendues plus efficaces en utilisant des méthodes telles que l'appariement en fréquence, le recrutement randomisé, l'échantillonnage stratifié ou l'échantil-

Material may be protected by copyright law (Title 17, U.S. Code)

lonnage à deux niveaux. Ces méthodes ont en commun deux caractéristiques: (1) les variables dites "de premier niveau" sont recueillies pour tous les sujets, tandis qu'un recueil complet des variables n'est fait que pour un sous-échantillon et (2) la sélection des sujets de ce sous-échantillon en vue de recueillir les variables "de deuxième niveau" dépend à la fois de l'existence de la maladie et des variables de premier niveau. Comme ces variables influent sur la constitution des sous-échantillons, les méthodes d'analyse standards imposent un modèle multiplicatif des effets combinés d'une variable de premier niveau et d'une variable de second niveau. Nous montrons qu'en utilisant des méthodes pour données manquantes on peut obtenir des estimateurs du maximum de vraisemblance pour les paramètres d'intérêt, y compris ceux caractérisant les interactions entre variables de premier et de deuxième niveaux. Les effets combinés peuvent alors être modélisés de manière flexible, avec prise en compte simultanée de modèles additifs et multiplicatifs. L'application de ces méthodes aux données préliminaires d'une étude cas-témoin de la relation entre le cancer du poumon et le sexe, l'âge et la consommation de tabac suggère que l'âge et la consommation de tabac ont des effets multiplicatifs.

References

Anderson, J. A. (1979). Robust inference using logistic models. *International Statistics Institute Bulletin* **48,** 35–53.

Baker, R. J. and Nelder, J. A. (1978). *The GLIM System: Release 3*. Oxford: Numerical Algorithms Group.

Breslow, N. E. and Cain, K. C. (1988). Logistic regression for two-stage case–control data. *Biometrika* **75,** 11–20.

Breslow, N. E. and Day, N. E. (1980). *Statistical Methods in Cancer Research. Volume I: The Analysis of Case–Control Studies*. Lyon: International Agency for Research on Cancer.

Breslow, N. E. and Zhao, L. P. (1988). Logistic regression for stratified case–control studies. *Biometrics* **44,** 891–899.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* **39,** 1–38.

Fears, T. R. and Brown, C. C. (1986). Logistic regression methods for retrospective case–control studies using complex sampling procedures. *Biometrics* **42,** 955–960.

Flanders, W. D. and Greenland, S. (1991). Analytic methods for two-stage case–control studies and other stratified designs. *Statistics in Medicine* **10,** 739–747.

Hsieh, D. A., Manski, C. F., and McFadden, D. (1985). Estimation of response probabilities from augmented retrospective observations. *Journal of the American Statistical Association* **80,** 651–662.

Kalbfleisch, J. D. and Lawless, J. F. (1988). Likelihood analysis of multi-state models for disease incidence and mortality. *Statistics in Medicine* **7,** 149–160.

Kleinbaum, D. G., Kupper, L. L., and Morgenstern, H. (1982). *Epidemiologic Research: Principles and Quantitative Methods*. Belmont, California: Lifetime Learning Publications.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. New York: Wiley.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. New York: Chapman and Hall.

Scott, A. J. and Wild, C. J. (1991). Fitting logistic regression models in stratified case–control studies. *Biometrics* **47,** 497–510.

Siemiatycki, J. and Thomas, D. (1982). Biological models and statistical interactions: An example from multi-stage carcinogenesis. *International Journal of Epidemiology* **10,** 383–387.

Thomas, D. C. and Greenland, S. (1985). The efficiency of matching in case–control studies of risk factor interactions. *Journal of Chronic Diseases* **38,** 569–574.

Wacholder, S. (1986). Binomial regression in GLIM: Estimating risk ratios and risk differences. *American Journal of Epidemiology* **123,** 174–184.

Wacholder, S. (1992). Selection of controls in case–control studies. II. Types of controls. *American Journal of Epidemiology* **135,** 1029–1041.

Walker, A. M. (1982). Anamorphic analysis: Sampling and estimation for covariate effects when both exposure and disease are known. *Biometrics* **38,** 1025–1032.

Weinberg, C. R. and Sandler, D. P. (1991). Randomized recruitment in case–control studies. *American Journal of Epidemiology* **134,** 421–432.

Weinberg, C. R. and Wacholder, S. (1990). The design and analysis of case–control studies with biased sampling. *Biometrics* **46,** 963–975.

Weinberg, C. R. and Wacholder, S. (1993). Prospective analysis of case–control data under general multiplicative-intercept risk models. *Biometrika* **80,** 461–465.

White, J. E. (1982). A two-stage design for the study of the relationship between a rare exposure and a rare disease. *American Journal of Epidemiology* **115,** 119–128.

Received September 1991; *revised August* 1992 *and February* 1993; *accepted February* 1993.

Material may be protected by copyright law (Title 17, U.S. Code)